Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(5), 1117-1122 | ISSN:2251-6727

Identification of Population Specific Biomarkers for the Early Detection and Treatment of Lung Cancer.

Binisha.R.B [1], Jeyakumar Natarajan [2], Shanmughavel Piramanayagam [1,*].

Affiliation 1; Computational Biology lab, Department of Bioinformatics, Bharathiar University
Affiliation 2;Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar

(Received: 11 June 2024 Revised: 16 July 2024 Accepted: 10 August 2024)

KEYWORDS

Lung Cancer

Biomarker identification

Differential Expression analysis

R Programming

RNA-Seq

Hub Genes

ABSTRACT:

Introduction: Lung cancer is a leading cause of cancer-related mortality, particularly in urban populations. Despite significant advances in research, understanding the genetic differences in lung cancer across diverse populations remains limited.

Objectives: This study aims to investigate differential gene expression (DEGs) in lung cancer among three populations: Asian, White, and African, to identify common upregulated genes and explore their potential as therapeutic targets.

Methods: Lung cancer and normal tissue samples from three populations (15 cancer samples and 6 normal tissue samples per group) were sourced from the TCGA-NCI portal and processed through the GDC portal. DESeq2 and ggplot2 packages in R Studio were used for DEG analysis and visualization. Only upregulated genes were considered for further analysis, and a Venn diagram was used to identify common DEGs. Protein-Protein Interaction (PPI) networks and hub genes were then identified, followed by Gene Ontology (GO) analysis of hub genes.

Results: A total of 1,303 common DEGs were identified across the three populations, with 15 hub genes found to be central to tumorigenic processes. PPI analysis highlighted key hub genes that may serve as potential therapeutic targets.

Conclusions: This study enhances our understanding of lung cancer biology across different populations and underscores the importance of common hub genes in tumorigenesis. These findings support personalized oncology approaches to improve outcomes and address health disparities.

1. Introduction

Lung cancer is one of the most frequently causing malignant tumors in the world, and it has become the leading cause of mortality in malignant tumors in urban populations [1]. Every year, almost 1.8 million new instances of lung cancer are diagnosed, along with 1.6 million deaths. Smokers are more likely to develop lung adenocarcinoma (LUAD), the most common subtype of lung cancer. However, in recent years, non-smoker morbidity from LUAD has increased significantly due to lack of effective diagnostic tools, LUAD is frequently detected at a later stage[2]. More than 60% of LUAD patients were identified to contain targetable gene abnormalities [3]. The high death rate and poor prognosis of lung cancer are mostly related to the difficulty of early detection. If patients were diagnosed early, the average 5-year survival rate might reach 85%. Tumor markers have become a frequent method of cancer diagnosis thanks to advances in molecular biology [4].

Bioinformatics analysis is widely used in cancer research to uncover genetic variations related with cancer[5]. High-throughput gene microarrays and RNASeq can identify differentially expressed genes (DEGs) between normal and tumor samples in humans and animals. This allows for further exploration of tumor molecular changes at various levels, including DNA, RNA, proteins, epigenetics, and metabolism [5]. Also, the biomarkers that are identified so far are not population specific. The term Population genemics was applied in this study, which implies individuals are probed with the goal of documenting common genetic variants and identifying how they are distributed among people within populations and among populations from different areas of the world [6]. Current studies have discovered that the mortality rate of lung

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(5), 1117-1122 | ISSN:2251-6727

cancer is connected to the time of diagnosis, indicating that earlier treatment might significantly improve patient survival if cancer is detected in its early stages. Numerous studies have attempted to uncover biomarkers that can be used to determine whether a patient has cancer by comparing the physiological state of patients and healthy people [7].

This study contributes to our understanding of the molecular mechanisms underlying lung cancer across diverse ethnic populations, shedding light on potential therapeutic targets and biomarkers with implications for precision medicine approaches. By delineating population-specific gene expression signatures, our findings may inform personalized treatment strategies tailored to the unique genetic makeup of individuals from different ethnic backgrounds, ultimately improving patient outcomes and reducing health disparities in lung cancer care.

2. Objectives

The objective of this study is to conduct a comprehensive analysis of differential gene expression (DEGs) in lung cancer across three distinct populations—Asian, White, and African—using data sourced from the TCGA-NCI and GDC portals. By analysing DEGs separately for each population and focusing exclusively on upregulated genes, the study aims to identify common genetic alterations involved in lung cancer tumorigenesis. Through the integration of DEGs via Venn diagram analysis, the study seeks to pinpoint shared upregulated genes across these populations. Furthermore, by constructing Protein-Protein Interaction (PPI) networks and identifying hub genes, the study aims to elucidate the molecular mechanisms underlying lung cancer progression and assess the biological significance of key genes in tumorigenic pathways. The ultimate goal is to contribute to a deeper understanding of the molecular drivers of lung cancer across diverse ethnicities, with a focus on uncovering potential therapeutic targets that may lead to the development of personalized oncology strategies aimed at reducing mortality and addressing health disparities in global populations.

3. Methods

DATA COLLECTION

The mRNA expression data of Lung cancer for Three population such as Asian, White and Africa were collected from TCGA-NCI portal [8] and downloaded using GDC Portal [9]. A total of 15 Lung cancer samples and six Normal tissue samples were retrieved for each population. The sample IDs were converted into Gene Symbols using DAVID tool.

DATA PREPROCESSING

The DEGs of the mRNA datasets for the three population were identified using the Statistical Software R Studio. For Performing R, the Data should be preprocessed by removing the NA values and rounding the integers into whole numbers. The packages such as “DESeq2”, “ggplot2”, were used for DEG analysis and plotting graph respectively.

IDENTIFICATION OF DEGs

Using the R package “DESeq2” (installed using Bioconductor), the DEGs were analyzed for White, Africa and Asian population separately. The genes that met the cutoff criteria such as P value <0.05 and $|\text{logfc}| \geq 1.0$ were considered as DEG. Genes with P value <0.05 and $|\text{logfc}| \geq 1.0$ is considered as Upregulated P value <0.05 and $|\text{logfc}| \leq -1.0$. The “ggplot2” package was used to plot the relative log Expression plot for the datasets.

INTEGRATION OF DEGs

The DEGs for the three population (White, Africa and Asian) were identified separately and only the upregulated genes were filtered for this study. The common genes from the three population were identified by showing Venn Diagram.

CONSTRUCTION OF PPI NETWORK AND HUB GENE IDENTIFICATION

The Search Tool for the Retrieval of Interacting Genes (STRING; http://string.embl.de/) is a biological database meant to build a PPI network of DEGs based on known and projected PPIs, and then investigate the functional connections between proteins. The Hub genes were identified using CytoHubba plugin in cytoscape (https://cytoscape.org/) with Degree ranking greater than 65 for better analysis.

GO ENTRICHMENT ANALYSIS FOR HUB GENES

GO enrichment analysis is mostly used to annotate genes and gene products. The GO is divided into three categories: biological processes (BP), cellular composition (CC), and molecular function (MF). The Database used is ShinyGO 0.80 and the statistical significance was adjusted as P value is 0.05.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(5), 1117-1122 | ISSN:2251-6727

4. Results

IDENTIFICATION AND INTEGRATION OF DEGs

Differential Expression analysis identified 1316 upregulated DEGs for African and Asian population and 1303 upregulated DEGs for White (Figure 1). We found 1302 genes were common for all the population and 14 genes were common for Asian and African population (Figure 1). Table 1 shows the list of genes that are common for the populations.

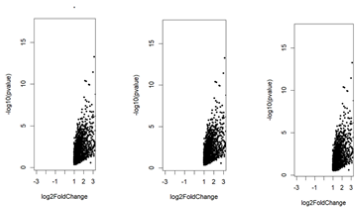

Figure 1: Volcano graph showing the upregulated (log2FC Value above 1) DEGs of Asian, African and White population respectively. Also, the venn Diagram showing common number of genes in the population.

PROTEIN-PROTEIN INTERACTION and HUB GENE IDENTIFICATION

In total, 1303 (Common genes) were taken for the STRING interaction analysis and identified the average node length is 12 with 1299 nodes and 2289 edges, having PPI enrichment p-value is greater than 1.0e-16. Next, the PPI network was constructed using Cytoscape. The most important genes (Hub genes) were identified using cytohubba plugin and top 15 genes having degree above 50 were selected as hub genes. The genes are ACTB, RPS2, CCNA2, CDC20, CCND1, CDH1, RPS16, AURKB, H2AZ1,H2AX, UBE2C, CYCS, HDAC1, CDC45, KIF23 (Figure 2).

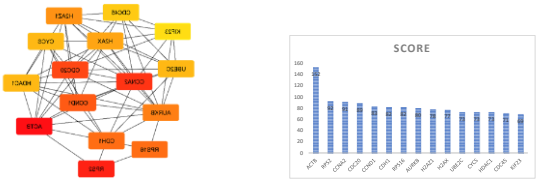

Figure 2: Showing the 15 hub genes (Red shaded with higher degree and yellow shaded with lower degree) and the graph showing the Genes with degree score respectively.

GO ENTRICHMENT ANALYSIS FOR HUB GENES

The Function of all the hub genes and their impact on cancer has been studied using the GO analysis (Table 2).

Table 2: The Functions of Selected Hub genes Symbol

Symbol	Description
HDAC1	**histone deacetylase 1** (HDAC1 upregulation can lead to the silencing of tumor suppressor genes by removing acetyl groups from histones, resulting in a more condensed chromatin structure and reduced gene expression)
CDC20	**cell division cycle 20** (CDC20 activates the anaphase-promoting complex/cyclosome (APC/C), an E3 ubiquitin ligase that targets specific cell cycle proteins for degradation. This activity is crucial for the proper segregation of chromosomes during mitosis.)
H2AZ1	**H2A.Z variant histone 1** (is a variant of the histone H2A protein family, which plays a crucial role in the structural organization of chromatin and the regulation of gene expression.)
CCNA2	**cyclin A2** (Cyclin A2 is a regulatory protein involved in the cell cycle, playing a crucial role in the transition from the G1 phase to the S phase and from the G2 phase to the M phase. Its overexpression can disrupt normal cell cycle control, leading to uncontrolled cell proliferation, a hallmark of cancer)
ACTB	**actin beta** (Actin beta is one of the isoforms of actin, a fundamental component of the cytoskeleton involved in various cellular functions, including maintenance of cell shape, motility, and division.)
CYCS	**cytochrome c, somatic**(cytochrome c is released from mitochondria into the cytoplasm, where it binds with Apaf-1 (apoptotic protease activating factor-1) to form the apoptosome, initiating the

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(5), 1117-1122 | ISSN:2251-6727

	caspase cascade and ultimately leading to programmed cell death.)
CCND1	**cyclin D1** (Cyclin D1 is a protein involved in regulating the cell cycle progression from G1 to S phase by binding to and activating cyclin-dependent kinase 4 (CDK4) or CDK6.)
H2AX	**H2A.X variant histone** (H2A.X is a histone variant involved in DNA damage response and repair. When DNA damage occurs, H2A.X becomes phosphorylated, marking the site of damage and facilitating repair processes.)
KIF23	**kinesin family member 23**(KIF23 is a member of the kinesin superfamily of proteins, which are motor proteins involved in various cellular processes, including cell division (mitosis) and intracellular transport.)
RPS2	**ribosomal protein S2** (RPS2 is a component of the ribosome, which is responsible for protein synthesis in cells)
CDH1	**cadherin 1** (Cadherin 1, also known as E-cadherin, is a calcium-dependent cell adhesion molecule that plays a crucial role in maintaining tissue integrity and cell-cell adhesion in normal epithelial tissues)
AURKB	**aurora kinase B** (Aurora kinases are a family of serine/threonine kinases that play crucial roles in cell division, particularly in regulating mitosis.)
RPS16	**ribosomal protein S16** (Ribosomal proteins like RPS16 are essential components of the ribosome, the cellular machinery responsible for protein synthesis, including cell proliferation, differentiation, and apoptosis.)
UBE2C	**ubiquitin conjugating enzyme E2 C** (The UPS is responsible for targeted protein degradation and regulates various cellular processes, including cell cycle progression, DNA repair, and apoptosis.)
CDC45	**cell division cycle 45** (CDC45 is a key protein involved in DNA replication and is essential for the initiation and progression of the cell cycle. When CDC45 is dysregulated and overexpressed, it can lead to aberrant cell proliferation, genomic instability, and ultimately contribute to carcinogenesis.)

5. Discussion

This study analyzed gene expression data from lung cancer patients across Asian, African, and White populations. The reference is driven from the previous study for prostate cancer, which emphasise only for two populations [10]. This work identified differentially expressed genes (DEGs) specific to each population and a core set of genes commonly dysregulated in all three groups. The identification of unique DEGs in each population highlights the genetic diversity of lung cancer. This suggests that factors like ethnicity might influence susceptibility to certain gene mutations. Also, by understanding these population-specific variations could be crucial for developing personalized treatment strategies. The future research can delve deeper into the functional roles of these population-specific DEGs to understand their contribution to lung cancer development and progression in each group.

The past studies suggest that CDC20 and CDC45 Plays a major role in lung cancer for both smokers and non-smokers [11]. AURKB is an important gene which triggers various cancers like lung cancer, Gastric cancer and Renal cancer. But it is unclear that this gene has an impact in different populations [12]. The other genes are also identified as biomarkers in epigenetic and expression profiles. These genes are proven as biomarkers irrespective of population. But this study discovers a set of genes commonly altered across all populations points towards core molecular pathways driving lung cancer, irrespective of ethnicity . These genes might represent promising targets for universal biomarkers for early detection or therapeutic interventions effective across diverse populations. Further investigation into the mechanisms by which these common DEGs contribute to lung cancer is necessary to validate them as potential targets.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(5), 1117-1122 | ISSN:2251-6727

The recent study, focussed on DEG Analysis using GEOR2 for microarray datasets and the hub genes were identified and further q-PCR technique was implemented for confirmation [13]. Our study focused on gene expression analysis of latest samples from TCGA database and R package was employed for pre-processing the samples and to identify the DEGs. Also, our future studies could incorporate additional data layers, like protein expression or mutational analysis, to provide a more comprehensive picture of lung cancer biology for all populations. The functional roles of the identified genes are also notified in the study and need further investigation to validate their potential as biomarkers or therapeutic targets. Overall, this study provides valuable insights into the genetic landscape of lung cancer across populations. Also, these genes are not specific for Indians, it can be used as a biomarker for White people and the Africans. The identification of population-specific variations, commonly dysregulated genes, and hub genes paves the way for developing more targeted diagnostic and therapeutic approaches for lung cancer patients across ethnicities.

6. Conclusion

In this study, the analysis of mRNA expression data from lung cancer samples across Asian, African, and White populations has unveiled a rich landscape of differentially expressed genes (DEGs). Through meticulous data preprocessing and sophisticated differential expression analysis techniques employing tools like R Studio and DESeq2, we have identified a considerable number of upregulated DEGs unique to each population, reflecting the inherent genetic diversity and complexity of lung cancer across populations.

Moreover, our integration of DEGs has uncovered a noteworthy subset of genes that are commonly dysregulated across all three populations. This discovery of shared genetic alterations underscores the existence of fundamental molecular pathways driving lung cancer pathogenesis, transcending ethnic boundaries. The identification of these common DEGs opens avenues for the development of universal biomarkers and therapeutic targets with broad applicability in lung cancer management.

Furthermore, our exploration of protein-protein interaction (PPI) networks has elucidated key hub genes that serve as central regulators in the intricate molecular networks underlying lung cancer. These hub genes, including ACTB, RPS2, CCNA2, CDC20, and others, emerge as critical players orchestrating various cellular processes implicated in lung cancer progression. Their centrality within the PPI network highlights their potential as pivotal nodes for targeted therapeutic interventions aimed at disrupting tumorigenic pathways and impeding cancer growth and metastasis.

In conclusion, this comprehensive analysis contributes significantly to our understanding of the molecular mechanisms driving lung cancer across diverse populations. By unraveling common genetic signatures and pivotal hub genes, this study not only provides valuable insights into the pathobiology of lung cancer but also lays the groundwork for the development of precision oncology strategies tailored to individual patients, regardless of their ethnic backgrounds. Moving forward, further investigation into the functional significance of identified DEGs and hub genes holds promise for the advancement of personalized approaches to lung cancer diagnosis, prognosis, and treatment, ultimately improving patient outcomes and advancing the field of oncology.

Refrences

[1] Zhang, L.; Peng, R.; Sun, Y.; Wang, J.; Chong, X.; Zhang, Z. Identification of Key Genes in Non-Small Cell Lung Cancer by Bioinformatics Analysis. *PeerJ*, **2019**, *7*, e8215. https://doi.org/10.7717/peerj.8215.

[2] Wu, Y.; Yang, L.; Zhang, L.; Zheng, X.; Xu, H.; Wang, K.; Weng, X. Identification of a Four-Gene Signature Associated with the Prognosis Prediction of Lung Adenocarcinoma Based on Integrated Bioinformatics Analysis. *Genes*, **2022**, *13* (2), 238. https://doi.org/10.3390/genes13020238.

[3] Guo, T.; Ma, H.; Zhou, Y. Bioinformatics Analysis of Microarray Data to Identify the Candidate Biomarkers of Lung Adenocarcinoma. *PeerJ*, **2019**, *7*, e7313. https://doi.org/10.7717/peerj.7313.

[4] Chen, Y.; Ma, Z.; Min, L.; Li, H.; Wang, B.; Zhong, J.; Dai, L. Biomarker Identification and Pathway Analysis by Serum Metabolomics of Lung Cancer. *BioMed Res. Int.*, **2015**, *2015*, e183624. https://doi.org/10.1155/2015/183624.

[5] Li, Z.; Sang, M.; Tian, Z.; Liu, Z.; Lv, J.; Zhang, F.; Shan, B. Identification of Key Biomarkers and

Potential Molecular Mechanisms in Lung Cancer by Bioinformatics Analysis. *Oncol. Lett.*, **2019**, *18* (5), 4429–4440. https://doi.org/10.3892/ol.2019.10796.

[6] Nedelkov, D.; Kiernan, U. A.; Niederkofler, E. E.; Tubbs, K. A.; Nelson, R. W. Population Proteomics: The Concept, Attributes, and Potential for Cancer Biomarker Research *. *Mol. Cell. Proteomics*, **2006**, *5* (10), 1811–1818. https://doi.org/10.1074/mcp.R600006-MCP200.

[7] Tu, Z.; He, X.; Zeng, L.; Meng, D.; Zhuang, R.; Zhao, J.; Dai, W. Exploration of Prognostic Biomarkers for Lung Adenocarcinoma Through Bioinformatics Analysis. *Front. Genet.*, **2021**, *12*. https://doi.org/10.3389/fgene.2021.647521.

[8] The Cancer Genome Atlas Program (TCGA) - NCI https://www.cancer.gov/ccg/research/genome-sequencing/tcga (accessed May 16, 2024).

[9] GDC Data Transfer Tool | NCI Genomic Data Commons https://gdc.cancer.gov/access-data/gdc-data-transfer-tool (accessed May 16, 2024).

[10] Agalliu, I.; Kwon, E. M.; Salinas, C. A.; Koopmeiners, J. S.; Ostrander, E. A.; Stanford, J. L. Genetic Variation in DNA Repair Genes and Prostate Cancer Risk: Results from a Population-Based Study. *Cancer Causes Control*, **2010**, *21* (2), 289–300. https://doi.org/10.1007/s10552-009-9461-5.

[11] He, X.; Zhang, C.; Shi, C.; Lu, Q. Meta-Analysis of mRNA Expression Profiles to Identify Differentially Expressed Genes in Lung Adenocarcinoma Tissue from Smokers and Non-Smokers. *Oncol. Rep.*, **2018**, *39* (3), 929–938. https://doi.org/10.3892/or.2018.6197.

[12] Wan, B.; Huang, Y.; Liu, B.; Lu, L.; Lv, C. AURKB: A Promising Biomarker in Clear Cell Renal Cell Carcinoma. *PeerJ*, **2019**, *7*, e7718. https://doi.org/10.7717/peerj.7718.

[13] Yadav, D. K.; Bhadresha, K.; Rao, P.; Shaikh, S.; Rawal, R. M. Identification of Hub Genes Associated with Prognosis of Lung Cancer via Integrated Bioinformatics and in Vitro Approach. *J. Biomol. Struct. Dyn.*, **2023**, *41* (20), 11204–11218. https://doi.org/10.1080/07391102.2022.2160816.